Licenciatura en Ciencias del Comportamiento

Ciencia de Datos

Trabajo Práctico N° 1: Un primer encuentro con la EPH

Autores

Facundo Tomás Villegas Leiva

Clara Schmukler

Sofía Abril Cortada Conti

Prof. Magistral

María Noelia Romero

Prof. Tutorial

Ignacio Anchorena

Semestre de Primavera 2025

Parte I: Familiarizandonos con la base EPH y limpieza

Link a repositorio de GitHub:
https://github.com/cschmukler/Ciencia-de-Datos---Grupo-11.git

1. El INDEC mide la pobreza con el método de línea de pobreza (LP) que consiste en ver si las personas pueden satisfacer con sus ingresos sus necesidades alimentarias y no alimentarias esenciales a través de la compra de bienes y servicios.
   El INDEC elabora una canasta básica que considera alimentos así como bienes y servicios no alimentarios y luego analiza la proporción de hogares que no pueden comprarla.
2.
   a. La región del país en específico que elegimos para trabajar en el trabajo fue la región Patagonia.
   b. De las 15 variables seleccionadas, se las cuales se encuentran las solicitadas: CH04 (sexo), CH06 (edad), CH07 (estado civil), CH08 (cobertura paga), NIVEL_ED (nivel educativo), ESTADO (condición de actividad), CAT_INAC (cantidad de tiempo inactivo) e IPCF (monto de ingreso per cápita familiar). También están las que elegimos también: V18_M (monto del ingreso por otros ingresos en efectivo como limosnas o juegos de azar), PP08D1 (monto total cobrado por sueldos, salario familiar, horas extras y otros complementos), CH09 (sabe leer o escribir), PP04C (cantidad de personas que trabajan en su empleo), CH12 (nivel más alto de cursada), PP05C_2 (si el negocio tiene local) y PP07C (si el empleo tiene tiempo de finalización).

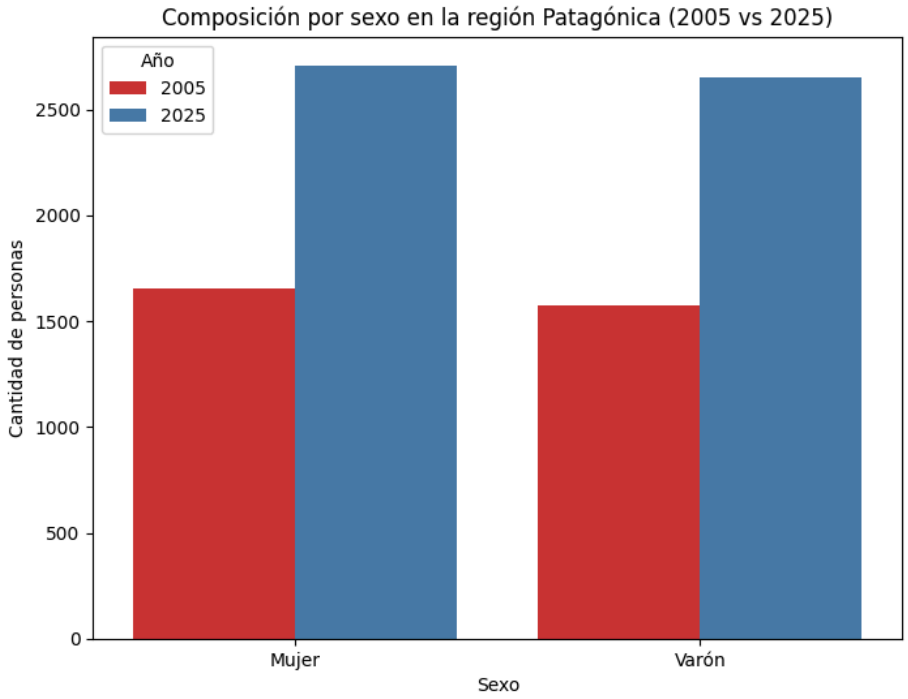
Al analizar el gráfico de valores faltantes por variable y año, observamos que en 2005 no se registran datos faltantes en ninguna de las 15 variables seleccionadas. En cambio, en el año 2025 aparecen valores faltantes en cuatro variables específicas: PP08D1, PP04C, PP05C_2 y PP07C, todas con 3036 valores faltantes. Por lo tanto, estas son las variables que tienen más valores faltantes, y aparecen solo en el año 2025.

c. Después de revisar todos los datos no borramos nada porque no vimos ninguna variable sin sentido, o sea, no encontramos valores realmente extraños. Lo que sí notamos es que en algunas preguntas faltaban muchas respuestas, pero cada una de esas faltas tiene una justificación. En V18_M, que se refiere al monto del ingreso por otros ingresos en efectivo como limosnas o juegos de azar, la mayoría aparece como 0.0 porque la gente directamente no tiene este tipo de ingresos. En CAT_INAC, que son las categorías de inactividad, muchos casos no están respondidos porque esas personas no son inactivas. En CH12, que pregunta por el mayor nivel de cursada, los 0.0 se pueden interpretar como que no tienen ningún nivel de cursada. En PP05C_2, que pregunta si el lugar de trabajo tiene local en caso de trabajo independiente, los 0.0 corresponden a que no son trabajadores independientes. Y en PP07C, que se dirige a los trabajadores asalariados y pregunta si el empleo tiene tiempo de finalización, los 0.0 se pueden interpretar como personas que no son asalariadas.

Parte II: Primer Análisis Exploratorio

3. Comenzando con el primer análisis exploratorio, elaboramos un gráfico de barras comparativo de la variable CH04 (sexo) para los años 2005 y 2025 en la región Patagónica (Figura 2). Podemos observar que en 2005 la distribución entre varones y mujeres era relativamente equilibrada, con una ligera mayor presencia femenina. En 2025 podemos ver que ambas categorías muestran un incremento en cantidad absoluta, reflejando el crecimiento de la población de la región Patagónica, además de conservar el equilibrio entre sexos. Podemos decir que, en términos de proporción, la relación entre mujeres y varones se mantuvo estable, lo que nos permite afirmar que no hubo cambios significativos en la distribución por sexo en la región a lo largo de los años. Por último, destacamos la diferencia de magnitud entre cantidad de personas registradas por la encuesta en 2005 y 2025. Este incremento tan notorio sugiere un aumento de la población encuestada y probablemente un crecimiento real de la población en la región.

Composición de la población por sexo (mujeres y varones) de la región Patagónica para los años 2005 y 2025. Se observa un aumento en la cantidad de personas en ambos grupos, manteniéndose una distribución equilibrada entre sexos en ambos años:

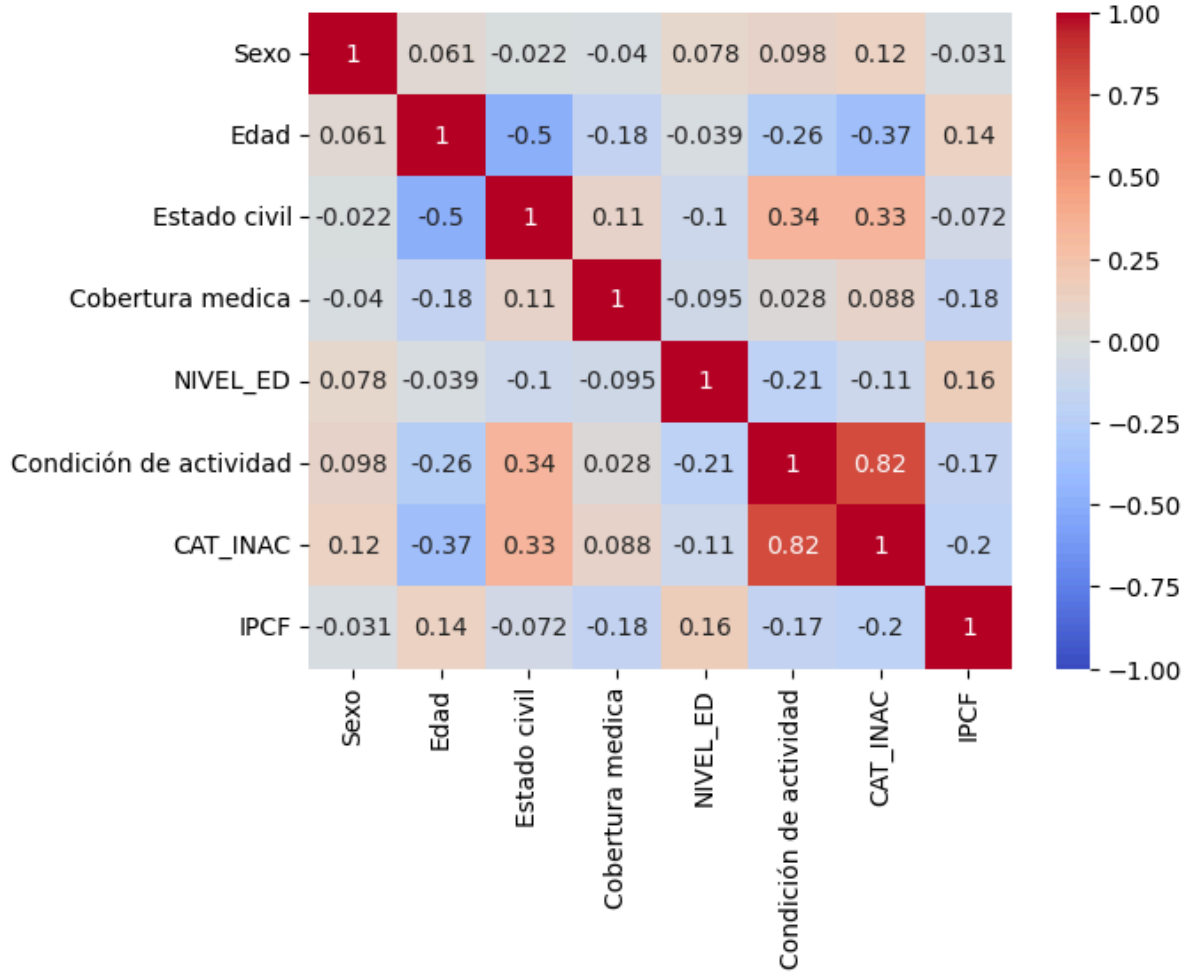


4. Para la matriz de correlación primero seleccionamos las variables de interés: CH04, CH06, CH07, CH08, NIVEL_ED, ESTADO, CAT_INAC e IPCF y creamos una nueva base para cada año con estas variables. Con las bases ya creadas, comenzamos a

crear las variables dicotómicas binarias necesarias para poder realizar la matriz de correlación. Además, renombramos las columnas para que tuvieran sentido en el gráfico y renombramos las variables utilizando el diccionario de cada año.

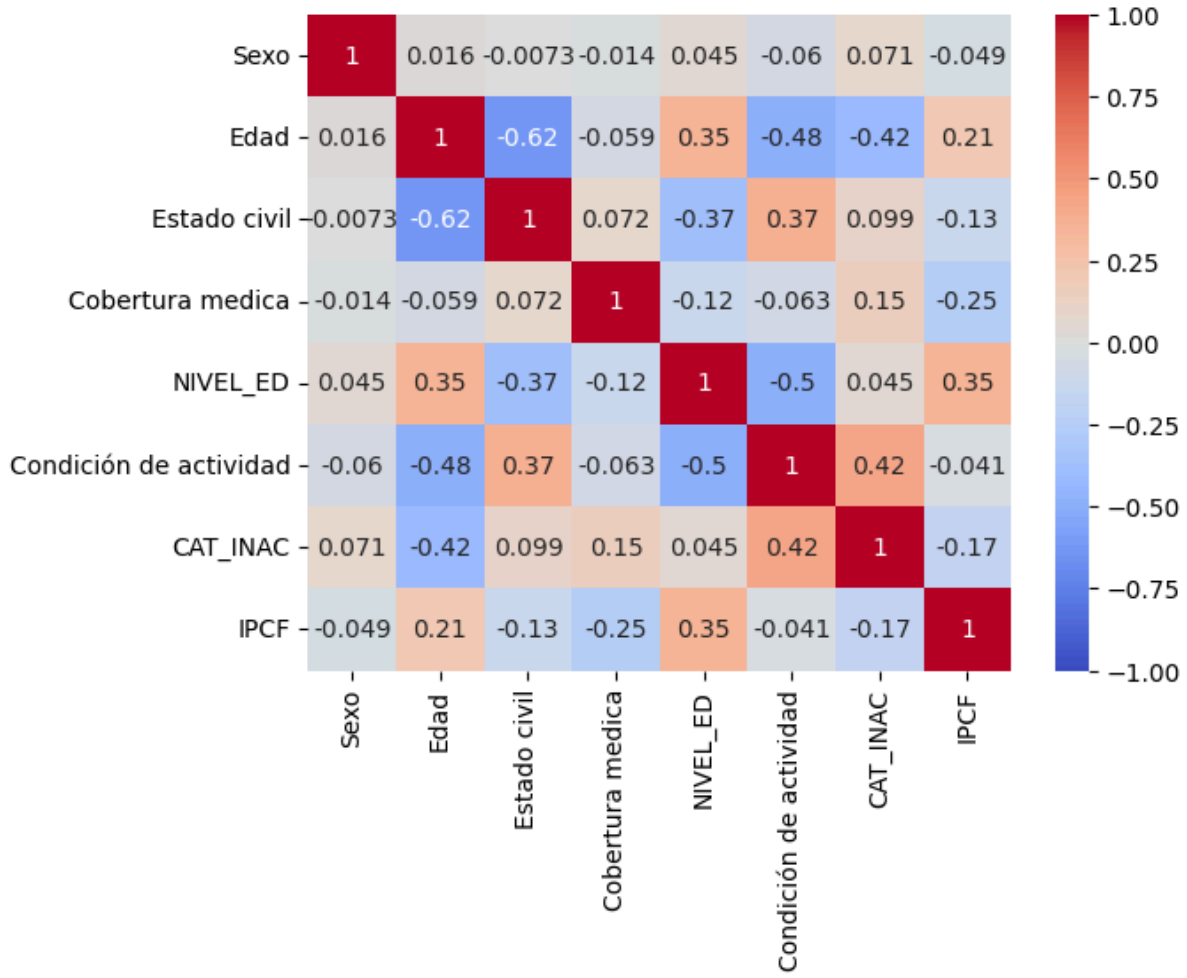
Creamos la matriz de correlación para EPH 2025 Patagónica y la visualizamos con un heatmap. Los valores van de -1 a +1, se muestran los números en cada celda y usamos la paleta "coolwarm", utilizando los comandos disponibles en Delft Stack. Hacemos lo mismo apra EPH 2005.

Matriz de correlación de variables seleccionadas en la EPH 2025 para la región Patagónica:



Se observa una correlación positiva fuerte entre la condición de actividad y la categoría de inactividad, mientras que la edad se asocia negativamente con el estado civil y con la condición de actividad. El ingreso per cápita familiar (IPCF) muestra correlaciones débiles con la mayoría de las variables.

Matriz de correlación de variables seleccionadas en la EPH 2005 para la región Patagónica:



La edad muestra una correlación negativa con el estado civil y la condición de actividad, mientras que el nivel educativo (NIVEL_ED) presenta asociaciones positivas con el IPCF y con la condición de actividad.

Parte III: Conociendo a los pobres y no pobres

5. En la variable ITF, que corresponde al monto de ingreso total familiar, observamos que 1225 personas no respondieron mientras que 7363 sí lo hicieron. Esto significa que aproximadamente un 14% de los casos aparece como no respondido.
6. En 2025 se identificaron 2247 pobres, lo que representa aproximadamente un 41,93% de la muestra, mientras que en 2005 se registraron 610 pobres, equivalentes al 18,89% de la muestra.

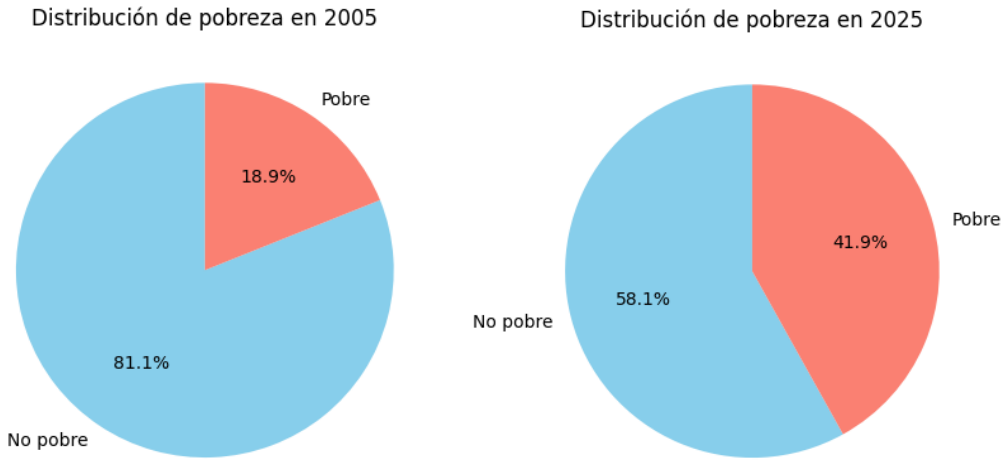
7. La variable pobre toma valor 1 si el hogar no alcanza a cubrir la Canasta Básica Total (CBT) y 0 en caso contrario.

   En 2005, sobre un total de 3.229 observaciones, el 18,9% de los hogares fueron identificados como pobres. En cambio, en 2025, sobre 5.359 observaciones, el porcentaje de hogares pobres asciende a 41,9%.

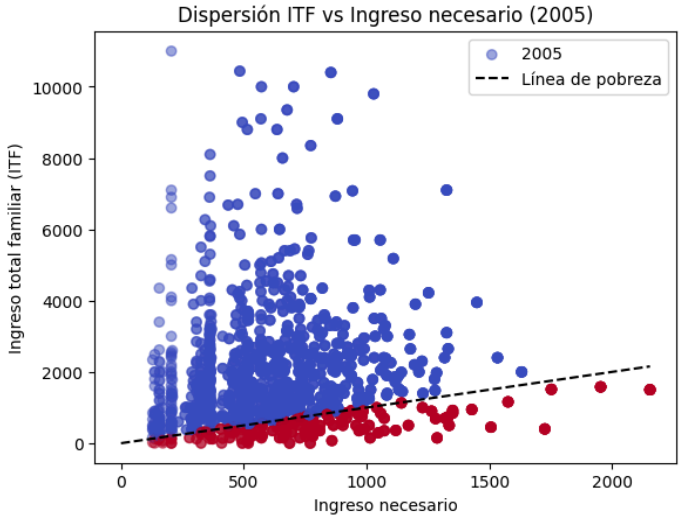
   La distribución confirma que en 2005 la mayoría de los hogares se encontraban por encima de la línea de pobreza (mediana = 0), mientras que en 2025 la pobreza adquiere mayor relevancia, con una dispersión más alta (desvío estándar de 0,49 frente a 0,39 en 2005).

```
Estadísticas descriptivas de la variable 'pobre':
              2005         2025
count  3229.000000  5359.000000
mean      0.188913     0.419295
std       0.391500     0.493490
min       0.000000     0.000000
25%       0.000000     0.000000
50%       0.000000     0.000000
75%       0.000000     1.000000
max       1.000000     1.000000
```

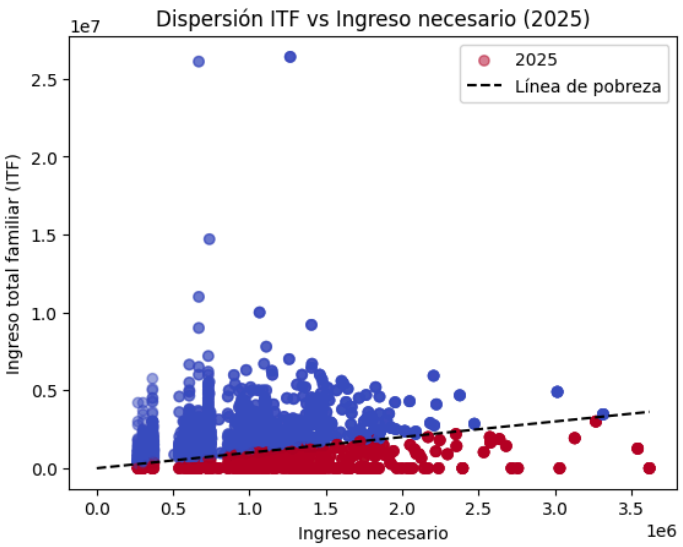
El gráfico muestra la distribución de la pobreza en 2005. Se observa que el 81,1% de los hogares no eran pobres, mientras que solo el 18,9% se encontraba por debajo de la línea de pobreza. Esto refleja que, en ese año, la gran mayoría de la población patagónica lograba cubrir el ingreso mínimo necesario para no ser pobre.



El gráfico muestra la distribución de la pobreza en 2025. En este caso, el porcentaje de hogares pobres asciende al 41,9%, mientras que los no pobres representan el 58,1%. Esto implica un incremento notable respecto de 2005, donde sólo el 18,9% de los hogares eran pobres. La comparación sugiere un empeoramiento significativo en las condiciones de vida, con una proporción de pobreza que se más que duplicó en veinte años.



El gráfico presenta la relación entre el ingreso total familiar (ITF) y el ingreso necesario para no ser pobre en 2005. Los puntos en azul representan hogares no pobres (ITF ≥ ingreso necesario), mientras que los puntos en rojo corresponden a los pobres (ITF < ingreso necesario). Se observa que la gran mayoría de los hogares se sitúan por encima de la línea de pobreza, lo que coincide con la baja proporción de pobreza (18,9%) encontrada para ese año. Sin embargo, también se identifican algunos hogares en situación de vulnerabilidad, con ingresos cercanos a la línea, lo cual muestra un riesgo potencial de caer en la pobreza ante cambios adversos en sus ingresos o en el costo de la canasta básica.



El gráfico muestra la dispersión entre el ingreso total familiar (ITF) y el ingreso necesario para no ser pobre en 2025. A diferencia de 2005, aquí se observa una mayor concentración de hogares por debajo de la línea de pobreza (puntos rojos), lo cual refleja el incremento de la pobreza hasta un 41,9% de la muestra. Además, los hogares pobres no solo son más numerosos, sino que presentan brechas más amplias respecto al ingreso necesario, indicando situaciones de pobreza más severa. En contraste, los hogares no pobres (puntos azules) se ubican con mayor dispersión y algunos muestran ingresos muy elevados, lo que evidencia un aumento de la desigualdad en la distribución del ingreso.